## Dan Sanchez

## BAN 502

### Module 3 Assignment 2

### Dr. Stephen Hill

### June 1st, 2020

library(tidyverse)  
library(MASS)  
library(caret)  
library(ROCR)

parole=read.csv("parole.csv")

parole = parole %>% mutate(male = as\_factor(as.character(male))) %>%  
mutate(male = fct\_recode(male,  
"Female" = "0",  
"Male" = "1"))  
  
parole = parole %>% mutate(race = as\_factor(as.character(race))) %>%  
mutate(race = fct\_recode(race,  
"White" = "1",  
"Non White" = "2"))  
  
parole = parole %>% mutate(state = as\_factor(as.character(state))) %>%  
mutate(state = fct\_recode(state,  
"Other" = "1",  
"Kentucky" = "2",  
"Louisiana" = "3",  
"Virginia" = "4"))  
  
parole = parole %>% mutate(crime = as\_factor(as.character(crime))) %>%  
mutate(crime = fct\_recode(crime,  
"Other" = "1",  
"Larceny" = "2",  
"Drug-Related" = "3",  
"Driving-Related" = "4"))  
  
parole = parole %>% mutate(multiple.offenses = as\_factor(as.character(multiple.offenses))) %>%  
mutate(multiple.offenses = fct\_recode(multiple.offenses,  
"Single Offense" = "0",  
"Multiple Offenses" = "1"))  
  
parole = parole %>% mutate(violator = as\_factor(as.character(violator))) %>%  
mutate(violator = fct\_recode(violator,  
"Not Violated" = "0",  
"Violated" = "1"))

**Task 1: Split the data into training and testing sets. Your training set should have 70% of the data. Use a random number (set.seed) of 12345.**

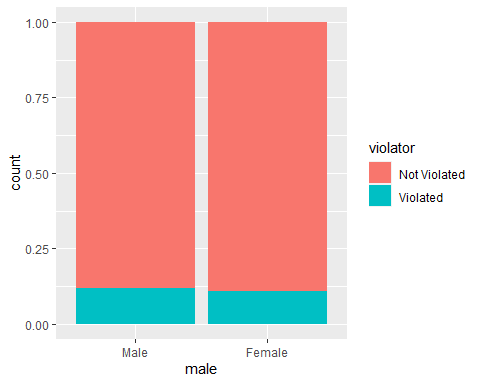
set.seed(12345)  
train.rows = createDataPartition(y=parole$violator, p=0.7, list=FALSE)  
train = slice(parole,train.rows)  
test = slice(parole,-train.rows)

**Task 2: Our objective is to predict whether or not a parolee will violate his/her parole. In this task, use appropriate data visualizations and/or tables to identify which variables in the training set appear to be most predictive of the response variable “violator”. Provide a brief explanation of your thought process.**

I have included data visualization and tables to represent how different variables impact the response variable “violator.” I found the tabular data more useful in pinpointing disparities between the variables. The three predictive variables with the greatest disparity were race, state, and multiple offenses. Other variables were relatively evenly split (male/female, and crime).

### Visualizations

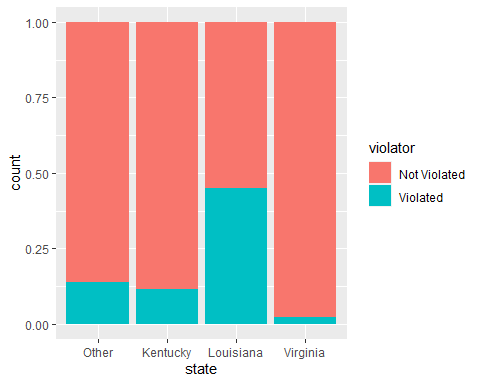
ggplot(parole, aes(x=male, fill=violator)) + geom\_bar(position="fill")



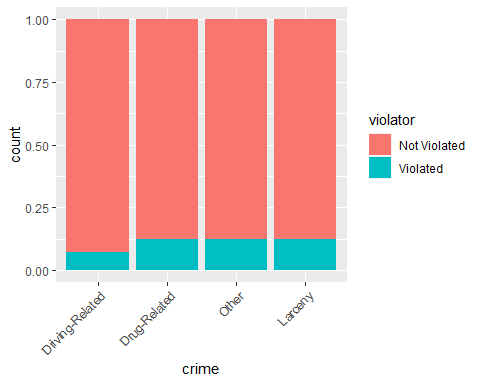
ggplot(parole, aes(x=race, fill=violator)) + geom\_bar(position="fill")



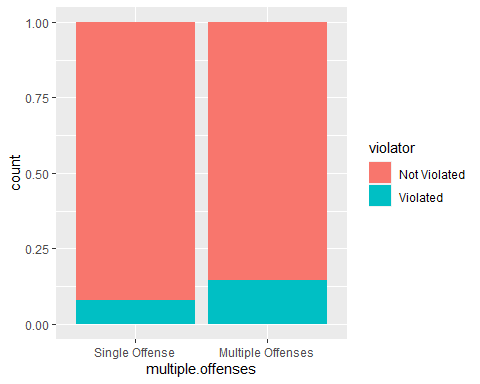
ggplot(parole, aes(x=state, fill=violator)) + geom\_bar(position="fill")



ggplot(parole, aes(x=crime, fill=violator)) + geom\_bar(position="fill") + theme(axis.text.x = element\_text(angle = 45, hjust = 1))



ggplot(parole, aes(x=multiple.offenses, fill=violator)) + geom\_bar(position="fill")



### Tabular Comparison

t1 = table(parole$violator,parole$male)  
t2 = table(parole$violator,parole$race)  
t3 = table(parole$violator,parole$state)  
t4 = table(parole$violator,parole$crime)  
t5 = table(parole$violator,parole$multiple.offenses)  
  
prop.table(t1,margin=2)

##   
## Male Female  
## Not Violated 0.8825688 0.8923077  
## Violated 0.1174312 0.1076923

prop.table(t2,margin=2)

##   
## White Non White  
## Not Violated 0.90488432 0.85664336  
## Violated 0.09511568 0.14335664

prop.table(t3,margin=2)

##   
## Other Kentucky Louisiana Virginia  
## Not Violated 0.86013986 0.88333333 0.54878049 0.97878788  
## Violated 0.13986014 0.11666667 0.45121951 0.02121212

prop.table(t4,margin=2)

##   
## Driving-Related Drug-Related Other Larceny  
## Not Violated 0.93069307 0.87581699 0.87619048 0.87735849  
## Violated 0.06930693 0.12418301 0.12380952 0.12264151

prop.table(t5,margin=2)

##   
## Single Offense Multiple Offenses  
## Not Violated 0.9201278 0.8535912  
## Violated 0.0798722 0.1464088

**Task 3: Identify the variable from Task 2 that appears to you to be most predictive of “violator”. Create a logistic regression model using this variable to predict violator. Comment on the quality of the model.**

mod1 = glm(violator ~ state, parole, family = "binomial")  
summary(mod1)

##   
## Call:  
## glm(formula = violator ~ state, family = "binomial", data = parole)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0955 -0.4981 -0.2071 -0.2071 2.7760   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.8165 0.2411 -7.534 4.92e-14 \*\*\*  
## stateKentucky -0.2079 0.3728 -0.558 0.577   
## stateLouisiana 1.6207 0.3277 4.946 7.58e-07 \*\*\*  
## stateVirginia -2.0153 0.4517 -4.461 8.15e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 483.27 on 674 degrees of freedom  
## Residual deviance: 382.89 on 671 degrees of freedom  
## AIC: 390.89  
##   
## Number of Fisher Scoring iterations: 6

Three factors in this model (Lousiana, Virginia, and Other) show significance with being predictive of “violator,” with P values less far below .05. The 390.89 AIC of this model was better (lower) than that of models examining race and multiple offenses.

**Task 4: Using forward stepwise, backward stepwise, or by manually building a model, create the best model you can to predict “violator”. Use only the training data set and use AIC to evaluate the “goodness” of the models. Comment on the quality of your final model. In particular, note which variables are significant and comment on how intuitive the model may (or may not) be.**

allmod = glm(violator ~., train, family = "binomial")   
summary(allmod)

##   
## Call:  
## glm(formula = violator ~ ., family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6055 -0.3932 -0.2643 -0.1384 2.9470   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.750397 1.318165 -2.845 0.00444 \*\*   
## maleFemale 0.137577 0.411340 0.334 0.73803   
## raceNon White 1.143719 0.403890 2.832 0.00463 \*\*   
## age 0.005279 0.016910 0.312 0.75490   
## stateKentucky 0.124282 0.492370 0.252 0.80072   
## stateLouisiana 0.217202 0.556154 0.391 0.69614   
## stateVirginia -3.801561 0.666733 -5.702 1.19e-08 \*\*\*  
## time.served -0.109344 0.118901 -0.920 0.35777   
## max.sentence 0.065956 0.054593 1.208 0.22700   
## multiple.offensesMultiple Offenses 1.711032 0.396463 4.316 1.59e-05 \*\*\*  
## crimeDrug-Related 0.516479 0.739095 0.699 0.48468   
## crimeOther 0.727043 0.690775 1.053 0.29257   
## crimeLarceny 1.119953 0.797552 1.404 0.16025   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 242.09 on 460 degrees of freedom  
## AIC: 268.09  
##   
## Number of Fisher Scoring iterations: 6

emptymod = glm(violator~1, train, family = "binomial")   
summary(emptymod)

##   
## Call:  
## glm(formula = violator ~ 1, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.4972 -0.4972 -0.4972 -0.4972 2.0745   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.0281 0.1434 -14.14 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 340.04 on 472 degrees of freedom  
## AIC: 342.04  
##   
## Number of Fisher Scoring iterations: 4

#forward stepwise  
forwardmod = stepAIC(emptymod, direction = "forward", scope=list(upper=allmod,lower=emptymod), trace = TRUE)

## Start: AIC=342.04  
## violator ~ 1  
##   
## Df Deviance AIC  
## + state 3 275.18 283.18  
## + max.sentence 1 331.01 335.01  
## + multiple.offenses 1 335.02 339.02  
## + race 1 336.51 340.51  
## + time.served 1 336.61 340.61  
## <none> 340.04 342.04  
## + crime 3 335.07 343.07  
## + male 1 339.72 343.72  
## + age 1 339.95 343.95  
##   
## Step: AIC=283.18  
## violator ~ state  
##   
## Df Deviance AIC  
## + multiple.offenses 1 254.96 264.96  
## + race 1 267.66 277.66  
## <none> 275.18 283.18  
## + max.sentence 1 274.27 284.27  
## + time.served 1 274.44 284.44  
## + age 1 275.11 285.11  
## + male 1 275.13 285.13  
## + crime 3 271.72 285.72  
##   
## Step: AIC=264.96  
## violator ~ state + multiple.offenses  
##   
## Df Deviance AIC  
## + race 1 246.98 258.98  
## <none> 254.96 264.96  
## + max.sentence 1 253.11 265.11  
## + time.served 1 254.47 266.47  
## + male 1 254.91 266.91  
## + age 1 254.94 266.94  
## + crime 3 252.75 268.75  
##   
## Step: AIC=258.98  
## violator ~ state + multiple.offenses + race  
##   
## Df Deviance AIC  
## <none> 246.98 258.98  
## + max.sentence 1 245.31 259.31  
## + time.served 1 246.33 260.33  
## + male 1 246.78 260.78  
## + age 1 246.98 260.98  
## + crime 3 244.78 262.79

summary(forwardmod)

##   
## Call:  
## glm(formula = violator ~ state + multiple.offenses + race, family = "binomial",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3609 -0.4094 -0.2705 -0.1575 2.9653   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.51087 0.36354 -6.907 4.96e-12 \*\*\*  
## stateKentucky 0.07372 0.46051 0.160 0.87282   
## stateLouisiana 0.10381 0.50018 0.208 0.83559   
## stateVirginia -3.60795 0.63788 -5.656 1.55e-08 \*\*\*  
## multiple.offensesMultiple Offenses 1.73482 0.39421 4.401 1.08e-05 \*\*\*  
## raceNon White 1.09382 0.38974 2.807 0.00501 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 340.04 on 472 degrees of freedom  
## Residual deviance: 246.98 on 467 degrees of freedom  
## AIC: 258.98  
##   
## Number of Fisher Scoring iterations: 6

Based on this data set, the best predictive model of the “violator” variable includes “state,” “race,” and “multiple.offenses.” These variables indicate significance, with p values less than .05. Including these variables makes sense, especially after considering the tabular comparison of variable performance in Task 2. The 258.98 AIC of this model is lower than that of any model examining the impact of a single variable. It is really interesting to note the “violator” spike for Louisiana, which makes me wonder if other data not included in this set might provide more insight.

**Task 5: Create a logistic regression model using the training set to predict “violator” using the variables: state, multiple.offenses, and race. Comment on the quality of this model. Be sure to note which variables are significant.**

mod2 = glm(violator ~ state + multiple.offenses + race, parole, family = "binomial")  
summary(mod2)

##   
## Call:  
## glm(formula = violator ~ state + multiple.offenses + race, family = "binomial",   
## data = parole)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4012 -0.4051 -0.2604 -0.1801 2.8739   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.50359 0.30055 -8.330 < 2e-16 \*\*\*  
## stateKentucky 0.04449 0.39449 0.113 0.9102   
## stateLouisiana 0.75016 0.39147 1.916 0.0553 .   
## stateVirginia -3.12945 0.51147 -6.119 9.44e-10 \*\*\*  
## multiple.offensesMultiple Offenses 1.51964 0.32027 4.745 2.09e-06 \*\*\*  
## raceNon White 0.74594 0.31828 2.344 0.0191 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 483.27 on 674 degrees of freedom  
## Residual deviance: 353.26 on 669 degrees of freedom  
## AIC: 365.26  
##   
## Number of Fisher Scoring iterations: 6

This logistic regression model shows signficance for “Multiple Offenses”, and “Non White.” “Louisiana” technically does not show significance with the established .05 threshold, but it is close at .0553. The AIC of 365.26 is in the neighborhood of single-variable models examined earlier in this exercise.

**Task 6: What is the predicted probability of parole violation of the two following parolees? Parolee1: Louisiana with multiple offenses and white race Parolee2: Kentucky with no multiple offenses and other race**

Parolee1 = data.frame(state="Louisiana", multiple.offenses="Multiple Offenses",race="White")  
predict(mod2, Parolee1,type="response")

## 1   
## 0.4418174

Parolee 1 would have a predicted probability of parole violation of .44.

Parolee2 = data.frame(state="Kentucky", multiple.offenses="Single Offense",race="Non White")  
predict(mod2, Parolee2,type="response")

## 1   
## 0.152755

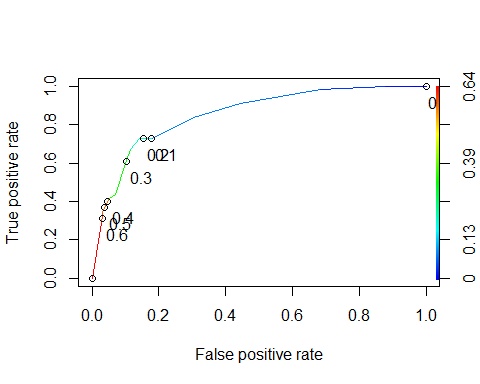
Parolee 2 would have a predicted probability of parole violation of .15.

**Task 7: Develop an ROC curve and determine the probability threshold that best balances specificity and sensitivity (on the training set).**

predictions = (predict(forwardmod, type="response"))  
head(predictions)

## 1 2 3 4 5 6   
## 0.07509978 0.19512504 0.19512504 0.07509978 0.07509978 0.19512504

ROCRpred = prediction(predictions, train$violator)   
ROCRperf = performance(ROCRpred, "tpr", "fpr")  
plot(ROCRperf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))



opt.cut = function(perf, pred){  
 cut.ind = mapply(FUN=function(x, y, p){  
 d = (x - 0)^2 + (y-1)^2  
 ind = which(d == min(d))  
 c(sensitivity = y[[ind]], specificity = 1-x[[ind]],   
 cutoff = p[[ind]])  
 }, perf@x.values, perf@y.values, pred@cutoffs)  
}  
print(opt.cut(ROCRperf, ROCRpred))

## [,1]  
## sensitivity 0.7272727  
## specificity 0.8588517  
## cutoff 0.2069629

The probability threshold that would best balance specificity and sensitivity is .2069629.

**Task 8: What is the accuracy, sensitivity, and specificity of the model on the training set given the cutoff from Task 7? What are the implications of incorrectly classifying a parolee?**

cmat = table(train$violator,predictions >0.6)  
(cmat[1,1]+cmat[2,2])/nrow(train)

## [1] 0.8921776

The accuracy of the model on the training set, given the cutoff of .2069629 is .8435518. The sensitivity is .7272727, and the specificity is .8588517.

Incorrectly classifying a parolee could misalign that individual’s parole terms with their perceived likelihoood of parole violation. Classifying a parolee as a likely violator may restrict their freedom and opportunties to violate parole. On the other hand, classifying a parolee as not likely to violate might provide too much flexibility in the terms, leading to a higher potential for parole violation.

**Task 9: Identify a probability threshold (via trial-and-error) that best maximizes accuracy on the training set.**

A probability threshold of 0.6 would maximize accuracy on the training set.

**Task 10: Use your probability threshold from Task 9 to determine accuracy of the model on the testing set.**

test\_preds = predict(forwardmod,newdata=test, type="response")

cmat = table(test$violator,test\_preds >0.6)  
(cmat[1,1]+cmat[2,2])/nrow(test)

## [1] 0.9108911

The accuracy of this model on the testing set comes to 91%.